DOCUMENT RESUME

ED 137 349 TM 006 150

AUTHOR Siracuse, Kathleen
TITLE Measuring the Achievement of Groups in Compensatory Education: An Alternative Testing Framework.
PUB DATE [Apr 77]
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; *Compensatory Education; *Criterion Referenced Tests; Diagnostic Tests; *Group Tests; *Item Banks; *Item Sampling; Language Programs; Norm Referenced Tests; Norms; Reading Programs; School Districts; Secondary Education; Secondary School Mathematics; Standardized Tests; Student Testing; Test Construction; Testing Problems; Testing Programs; Test Interpretation

ABSTRACT

An achievement testing framework is being developed by the Los Angeles Unified School District to assess the educational progress of 14,000 secondary level compensatory education students with something other than standardized tests. The technique of multiple matrix sampling was applied to the use of large item domains in the subject area of reading, mathematics, and language development. The domains of items were built locally on "content maps" which describe the skills actually taught in the compensatory education program. The process of constructing such frameworks is transportable to other programs. The possibility of obtaining normative data from the framework is being explored. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED137349

MEASURING THE ACHIEVEMENT OF GROUPS IN COMPENSATORY EDUCATION: AN ALTERNATIVE TESTING FRAMEWORK*

Author: Kathleen Siracuse
Los Angeles Unified School District
Research and Evaluation Branch
ESEA, Title I Programs

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 150

*Prepared for presentation at the annual meeting of American Educational Research Association, April 4-8, 1977, New York City

Interest in criterion-referenced testing (CRT) has accelerated in recent years, particularly as individualized approaches to learning have become more prevalent. This interest has lessened the emphasis placed on programs based only on broad educational goals. Concepts have now been added to criterion-referenced testing that suggest far-reaching, highly adaptable uses for group as well as individual assessment. These concepts as described by Millman (1972) include the establishment of large domains of items which collectively represent a proficiency standard in a subject area; and the construction of a matrix based on the total domain of items (or item pool). The matrix, when used horizontally, provides a set of either homogenous or heterogeneous items (in difficulty and format) that test a single objective. When used vertically, the matrix may be used to establish subtests of items which sample complete cross-sections of items across the entire domain of objectives for a subject area.

By making use of the subtests as detailed by Shoemaker (1974), and administering them in equal number to approximately equal portions of a student population at random, group scores are yielded which provide group diagnostic information on examinees' mastery of the objectives in a domain. As in the case of individual scores obtained by traditional uses of CRT's, the group is compared to a standard of achievement set by the criteria of the domain of objectives. The group is <u>not</u> compared to other groups on broad educational goals as in norm-referenced tests. (See also Shoemaker, 1975.)

<u>A New Framework for Achievement Testing</u>

By combining the concepts of criterion-referenced tests and multiple matrix sampling as described above, it has been possible to establish criteria for competency in the reading, mathematics and language development programs for the secondary level of compensatory education in Los Angeles, California. These criteria have served to define the domains of program objectives. Items were

purchased or generated that represented competency in meeting the objectives. The next step was to establish a matrix for each subject area as described, construct subtests and administer an experiemental testing framework that is being analyzed and revised to match the instructional program. The testing framework and the program are intended to become one and the same. The results will ultimately provide information on how well the actual objectives of the programs are being met, and suggest priorities for instructional decisions about the directions the overall programs should be taking.

The value of such results is that they offer group performance information, required for reporting to funding agencies of specially-funded programs. In addition to this, such group diagnostic scores serve to re-educate the community, the press, the parents, the educators and the source of funding about the uses of test scores. Since norm-referenced tests do not offer group diagnostic information, but only a comparison of groups to one another across diverse populations and on generalized skills, they are not useful in making competency-based educational decisions for priorities in program content. The much more specific information provided by group assessment with domain-referenced tests using multiple matrix sampling is far more enlightening, since it specifies what the group can and cannot do as the result of the objectives implemented in a program.

Normative Data From the Framework?

A further possible use of criterion-referenced tests for group assessments exists in the norming of the test results. As pointed out by Roudbush (1974), CTB/McGraw-Hill has already conducted research to determine the relationship between norm-referenced and criterion-referenced tests. Their initial findings were that well-written, comprehensive, criterion-referenced tests may be able to produce norm-referenced test results about as well as norm-referenced tests. This relationship

will be investigated in the present project as a component of interpreting results obtained from administering the framework.

Over a three-year period, framework test results will be compared to scores achieved by the same population on appropriate levels of the CTB/McGraw-Hill Comprehensive Tests of Basic Skills. Should the framework prove by this comparison to produce normative data, the way would then be open to eliminate redundant testing programs by using the framework for both criterion-referenced and normative data. This would help to meet the need for both types of evaluation that exists in specially-funded educational programs.

Uses of the Framework for Individual Diagnostic Purposes

The establishment of the matrix based on a large item pool also offers an opportunity for individual diagnostic measurements. As mentioned above, the matrix, when sliced horizontally contains many items covering a single objective in a subject area. Sets of diagnostic tests on single objectives or on small numbers of objectives may then be constructed. This increases the flexibility of the item pool or domain and provides educators with even more detailed assessments of student performances based on the same domain of objectives.

Issues of Test Security

By maintaining consistency between the program objectives and the evaluation criteria in all three forms of evaluation mentioned above, the need for traditional test security is eliminated. The program content and the test content become identical in what they require a student to demonstrate. Then, because the item pool is large (several hundred items) and because there are multiple forms of the test, it is no longer possible for the test to be memorized. A student may get any one of the various forms of the subtests during an examination.

Instructional Applications for the Framework

Teachers can use the testing framework's domain of objectives to plan program content. Results from pretests will show a class's and a grade level's performance on all the objectives. The idea will then be actually to teach to the framework's item format and difficulty, since it has been the program's objectives which determined the item pool content in the first place.

Training Teachers to Use the Framework

It will be necessary to conduct training sessions for teachers in the use of data from these domain-referenced achievement tests. This need for training was described by Shoemaker (1974) as being an essential component of converting to such a testing program. He states the following: "The critical ingredient here is creating a domain-referenced achievement testing atmosphere within the classroom and reorienting the teachers and students so that 'teaching specifically to the test'—or item domain, is perfectly acceptable and the primary goal of instruction." (p. 157) Test results from the framework are intended to provide group information specific enough about instructional needs that a teacher would be able to focus directly on skills and concepts in need of strengthening.

Flexibility of the Item Pools

An additional advantage of building a testing framework on an item pool or domain is that it allows an entire district (perhaps even a state) to unify itself around a pool of items which form a composite of their various programs. Individual schools (or districts) may then select those areas of the domain for which they wish to be held responsible. In other words, they may identify the portions of the domain they are actually trying to reach. When tests are administered across the entire domain, the results show how well their students are doing on areas that are not being formally taught in their particular program as well.

The ground has already been broken in this district for such assessment to take place on a broader scale. The Los Angeles Unified School District (LAUSD) is a large district (732,000 students) with a composite of specially-funded programs serving 142,248 students as of school year 1975-76. Approximately 14,000 of these students are involved in the secondary public schools portion of the programs. The state evaluation office for secondary ESEA Title I funded programs has granted LAUSD permission to compose its own version of a testing framework for language development. The framework is currently being built and is based on a language development curriculum written by educators in the secondary compensatory education program especially for the needs of the students involved. Again, the domain of objectives is the same for the program and the testing framework.

The acceptance of this type of data at the state level for specially-funded programs is a valuable precedent to have set. It is hoped that its use by the state to determine program effectiveness will serve to demonstrate the much more comprehensive nature of criterion-referenced test data for assessment of group performance as compared to the limited information obtained from norm-referenced standardized tests.

It is further hoped that the success of the framework would create an opening for the possibility of state-wide item pools and testing frameworks in subjects taught as components of specially-funded programs.

Applications of the Framework

The final product of this project is expected to be a testing framework for assessment of group performance in the subject areas of reading, mathematics and language development at the secondary level in compensatory education programs in Los Angeles. It is fully expected that the framework will be transportable to other districts conducting programs of a similar nature and with similar populations.

The actual process by which the framework is built can also be made transportable. A model for the construction of testing frameworks for group performance in any subjedt area and for any grade level will be developed. By using the model, districts anywhere in the country will be able to establish testing frameworks to match their programs. The actual format of the model has not yet been determined, but it may consist of such components as:

1) a written description of the steps to follow, methods to use, and types of personnel to involve, possible expenses, time tables and evaluation designs to use.

2) audio-visual aids to illustrate the above.

3) a list of available consultant services.

4) an annotated bibliography of resources in professional literature that relate to the construction and use of such frameworks.

Methods for the norming of comprehensive criterion-referenced test results will also be included in the process model if the attempt to carry this out becomes a successful component of the project.

Procedures for Constructing the Testing Framework

The process of designing and constructing the testing frameworks for assessment of group performance on item domains is already well under way for the Los Angeles secondary compensatory education programs.

Slightly different procedures were used for constructing each of the frameworks for reading, mathematics, and language arts, offering an opportunity to compare differences in approaches to some of the tasks involved. Basic tasks in each are:

1. Construction of Content Maps/Generation of Test Items
2. Scoring Procedures and Statistical Methods of Evaluating the Tests
3. Tryout and Use of the Tests
4. Revision/Refinement of Tests

Construction of Content Maps and Generation of Test Items

Alternative Approaches to Defining the Test Domain — The domains of objectives for the three frameworks were defined with three different approaches, although similar types of personnel were used in each case. The approaches, personnel, and working titles for the frameworks are described below.

Framework for Assessment in Reading (FAIR):

Series of workshops were held in November and December 1975 in which coordinators of reading programs at the school level and a reading teacher from each of three inner-city junior high schools were present. Also included were three reading-content advisors from administrative offices in the district, a district-hired content expert from Southwest Regional Laboratories, and an evaluator from

the research and evaluation branch for specially-funded programs in Los Angeles.

The objectives in the domain for this framework (or content map) were selected from a scope and sequence of the objectives for reading that exists as part of a reading management program used in the district. This program, entitled Developmental Reading Program (DRP), was written by the district and published through Paul Amidon and Associates in Minneapolis, Minnesota (Copyright 1972).

The portions of the scope and sequence that represented reading program content in the three junior high schools were selected to define what is meant by reading in the secondary compensatory education programs in this city. A total of 46 objectives were adopted covering a fairly wide range of skills. It was felt this would be necessary to accommodate the wide range of below-grade achievement levels in the programs while providing enough ceiling in the domain to discover what students already know about what may not have been taught formally.

Framework for Assessment in Mathematics (FAIM):

The workshops held to determine the domain of objectives in mathematics involved personnel in the same categories of positions as listed under FAIR (see above). The only difference was that all positions dealt with mathematics programs only. These workshops also took place in November and December 1975.

What did vary was the means by which the domain of objectives (or content map) was defined. In the FAIM workshop, its members generated the objectives for mathematics based on their experiences with what is actually taught and their knowledge of math content. They did not work

from a scope and sequence of objectives specified by a pre-existing packaged management system. The mathematics domain was defined to represent the range of skills and knowledges dealt with in the city's secondary compensatory education mathematics programs. As in FAIR, a certain amount of extra ceiling was added to the domain to detect serendipitous learnings. Sufficient floor was allowed as in FAIR to accommodate the range of below-grade achievement levels of students in these programs.

Framework for Assessment in English Skills (FAIES):

The domain of objectives (or content map) for language development skills and concepts was defined independently of the workshops for FAIR and FAIM. These workshops were held during summer 1975 for the purpose of defining and generating a curriculum management system for this subject area. The entire package was designed to meet the instructional needs of secondary compensatory education students in language development.

ESEA Title I personnel (English teachers, school program coordinators, and content advisors in language development) worked together to define and write a domain of objectives that would describe the language development program for the targeted student population. The group has subsequently produced the curriculum package in the form of a management system. The coordinators of the language development programs have been trained in the use of this system, and began implementing it in their school programs in the fall of this year as a field test of the materials.

The process of working from the domain of objectives for this system

to produce a matrix of test items organized into subtests has already been carried out. The selection of the items is discussed in the next subsection.

Item Generating Procedures — Test items for the three testing frameworks were acquired in three basic ways. They were purchased from item banks, used with copyright releases from publishers or generated by workshop members and by curriculum developers as in the case of items for FAIES. In all cases items were selected or revised to have four answer choices.

Framework for Assessment in Reading (FAIR):

Test items were selected from two sources of previously existing collections of items. These were the pre- and posttests for the Developmental Reading Program (DRP) mentioned earlier, and the National Assessment of Educational Progress released exercises published through the Superintendent of Documents, U. S. Government Printing Office, Washington, D.C., July 1973. Items contained in the DRP were generated by that program's developers.

The items from DRP were already coded to the objectives as part of the management system of that program. It was a simple matter, then, to locate and select test items to assess performance on objectives. In some cases, however, items were revised for greater relevancy to secondary level students. (The portion of the DRP used was originally written as a program for elementary students.)

Items taken from the National Assessment of Educational Progress materials were used only in the first experimental edition of FAIR. Since these items have been normed, they were included only for the purpose of comparison with similar items in the tests that came from DRP. They are not included in subsequent editions.

Where there were insufficient numbers of items available in the DRP materials or where existing items were inappropriately formated, workshop members generated test items based on the specifications of the objective and by drawing upon their knowledge of the content of reading.

Framework for Assessment in Mathematics (FAIM):

CO-OP items in mathematics were purchased from the University of Massachusetts. This item pool consisted of test items generated by public school and university personnel to cover the mathematics domain typical of "average" students in grades 4 - 9.

These items had full copyright releases on them and covered a wide range of skills and knowledges with large pools of items. Since the items were labeled to indicate the skill represented by them, workshop members were able to select those that matched the objectives in the domain for FAIM. It was discovered, however, that several of the items picked were improperly written. Corrections of errors were made by workshop members before the printing of the first experimental edition. In some cases, this correction necessitated the writing of a completely new item. The new items were modeled after the intent of the original items and formated similarly if appropriate.

Additional items were purchased from Instructional Objectives Exchange or IOX at UCLA. These also had full copyright releases. Not many of these items were used since few of them matched the domain of objectives developed for FAIM.

Framework for Assessment in English Skills (FAIES):

In the summer workshop for FAIES, curriculum management system materials developed by the workshop members for language development instruction

were available. These curriculum developers used the objectives they had written and their extensive knowledge of the content of the subject matter to select samples of items from the curriculum materials. These samples served as models for themselves and other writers to generate items for the objectives with some consistency. After the items were written, fellow workshop members critiqued the ability of the items to actually test the skill described in the objective.

Of the three methods used for selection of items for the framework, the one used for FAIR has proved most successful. The fact that the items were part of or based on a proven management system already field tested and in use in the district seems to have been beneficial. The use of the FAIR in field testing has produced the least amount of criticism of items for their appropriateness to the age and ability levels of students.

Still to come for all three portions of the framework, are workshops in which teachers and various content experts will critique test items for validity and for racial, ethnic and sexual bias.

Each set of tests will also be subjected to a critique by students for the purpose of gaining ideas on content for test items that would be of interest to the age level of the students taking them.

Critiques on the items in FAIM to date indicate that teachers feel many of the items are too difficult for students in the program. This raises an interesting issue, since it was teachers and coordinators from the same program who selected the items for FAIM. The conclusion need not be that this method of item selection is invalid, but simply that pitfalls are involved. Those selecting items may have a tendency to overestimate student achievement levels and provide items beyond the capability of the students.

Those administering the tests in reacting to the phenomenon of accountability, may have a tendency to underestimate student achievement, and suggest the elimination of items that may in fact be within the reach of a program's population.

The thing to be aware of here is that both tendencies occur, and can be reduced at least by cautioning those selecting the items about overestimating student abilities. Data obtained from field testing provides the information necessary for adjusting the difficulty of items during the revision process. A range of difficulty can then be provided to challenge but not overwhelm students. The idea is to have enough ceiling on the framework to measure growth in the achievement of the group, but also to provide enough bottom and middle range to comprehensively diagnose the group's performance on the actual heart of the program.

The method of item selection for FAIES has proved unsuccessful, but not because the actual method used is inherently bad. The selection of items from an unfield-tested collection of materials simply produced a set of tests that did not closely enough define the parameters of the language development programs being tested. Also, since the items used were actually drawn from much shorter pretests for a management package, their format did not lend itself to longer tests. As a result, it is necessary to conduct workshops involving various content experts (teachers, coordinators of reading programs, and district curriculum consultants) during which the content map (or domain of skills and concepts) for the language development component for compensatory education at the secondary level will be defined. Once these are defined, the existing items will have to be re-written or a new source of items located.

One of the crucial steps left out of the item selection process for FAIES

was the involvement of content experts who actually deal with the program students in deciding which skills and concepts are definitive of the program's objectives.

Test Construction Procedures — In the case of all three frameworks, subtests were constructed in the same manner. The items in the respective domains were arranged in a matrix as shown below. Items for a single objective were arranged in the row for that objective on the matrix. The rows of objectives represent the subdomains of single skills or concepts for the framework and for the program to be evaluated.

Items testing a single objective are arranged randomly in a row. The subtests are built by cutting the matrix vertically and using all the items in a column.

SAMPLE MATRIX*

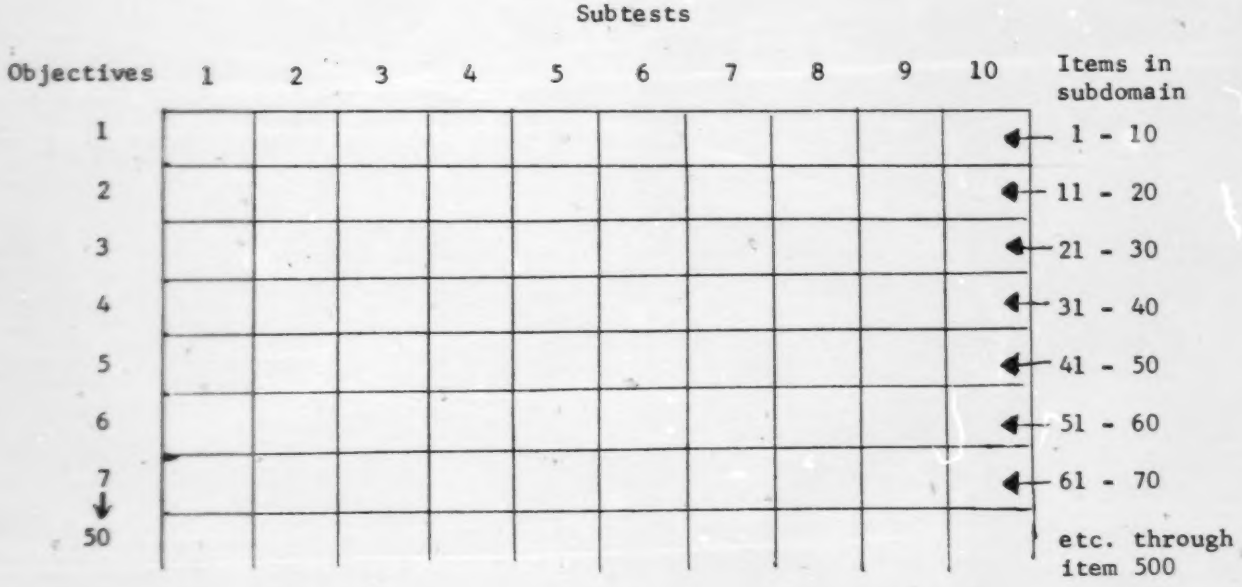

Figure 1

*For a hypothetical domain of 50 objectives having 500 items in the item pool.

Indexing Test Items to Curriculum for Diagnostic Purposes — In all three cases of the testing frameworks, the objectives for the test domain were given code numbers. Since the domains of objectives are descriptive of the respective programs involved, all test items are also coded to indicate the objective for which they are written. Then, by referring to these code numbers when reporting test results, it will be possible to report diagnostic information to teachers on students' abilities to perform on the objectives in the entire domain. (This relates to students' abilities to complete entire subtests. See the discussion of test length below.)

Scoring Procedures and Statistical Methods of Evaluating the Tests

Methods of Estimating Criterion Scores — As mentioned earlier, one of the primary goals of implementing the test frameworks is to be able to determine level of achievement of the secondary compensatory education students in Los Angeles over the three item domains for reading, mathematics and language development.

To accomplish this, the technique of multiple matrix sampling described earlier is being and will be used to administer both experimental and final editions of the tests. The subtests of items from the three domains are intended to be administered to randomly selected subgroups of students in grades 7 - 12. Although different subgroups take different subtests, it is pointed out by Shoemaker (1974) that the parameters estimated will be those that would have been obtained if all students had been tested on every item in the domain. It is, then, the ability of multiple matrix sampling to estimate achievement on large domains of items that makes it so valuable in assessing group achievement, since traditional norm-referenced tests only give this estimation on a very limited number of items. A further advantage

is that although a testing domain may consist of 500 items, each student takes a subtest of only 50 items. When all the subtest scores are combined, a composite score is obtained on each objective across the multiples of comparable items on all subtests.

As explained by Shoemaker (1974), "the results obtained from each subtest are used to estimate all parameters of interest." (p. 178) This means that on a single subdomain of items on an objective across all subtests, the results obtained on the same item type in each subtest are averaged or pooled to produce the single best estimate of that skill or concept. Standard errors of estimate are computed for each estimated skill or concept by using data obtained from all subtests.

The pooled estimate of a skill or concept is used to estimate the distribution of scores on each subdomain.

Establishing Cutoffs — The process of determining which areas of the domain should be completed with competency by any grade level in the programs is something that will evolve out of the interaction of teachers with test results. It was mentioned earlier that an advantage of working from a domain of objectives is that personnel operating a program can select those areas of the domain for which they feel it is reasonable to be held responsible.

It will be necessary for teachers and administrators to have experience with the three frameworks and the type of achievement data they provide before decisions can be made at the school level about criteria for levels of competency across the domain of objectives. Such competency criteria might demonstrate readiness for students to exit from compensatory education programs.

Normative data on grade equivalents for achievement within a subdomain or across the entire domain of objectives may help to establish such cutoffs.

This, of course, can only be done if it proves possible to obtain normative data on the testing framework. (See above discussion "Normative Data from the Framework?")

Determining Test Length — The first experimental editions of FAIR and FAIM were administered in the spring of 1976. At that time each of the subtests for FAIR were 51 items long. When test results were analyzed after this first field testing of the frameworks, it was found that significant numbers of students were unable to complete the latter portions of the subtests. Since FAIR and FAIM are domain-referenced tests, it is necessary that students have sufficient time to attempt all items on an entire subtest. As a result, the second experimental editions being administered in summer 1976 have been revised to reduce their length as follows:

Framework for Assessment in Reading (FAIR):

The items from the National Assessment of Educational Progress materials were eliminated. This reduced each subtest by five items, leaving 46 items in each. Since the items dropped were included only for comparison purposes, the actual item pool represented in FAIR was not reduced.

Framework for Assessment in Mathematics (FAIM):

It was felt that the length of time involved in solving the mathematics problems of FAIM subtests was at fault in students not completing all items. Not wishing to reduce the actual item pool, a solution was worked out whereby the ten original subtests were increased to fifteen subtests. This was done by distributing the 500 items throughout fifteen subtests in sequential order. In other words, the first fifteen items of the pool were distributed to each of the fifteen subtests. Subse-

quent items were similarly distributed. (See figure below.) This meant that each subtest no longer contained identical subdomains of items. Scoring for this change will be handled by adjusting the computer program used to analyze the data.

REVISED MATRIX FOR FAIM

Subtests

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	A	A	A	A	A	A	A	A	A	S	S	S	S	S
S	S	S	S	S	M	M	M	M	M	M	M	M	M	M
D	D	D	D	D	D	D	D	D	D	and so on ...				

A = Addition
S = Subtraction
M = Multiplication
D = Division
...and so on through 57 objectives

Figure 2

Framework for the Assessment of English Skills (FAIES):

The Framework for Assessment of English Skills was field tested in January 1977. The length of subtests for this first edition is 30 items. Indications are that format and difficulty of items make it impossible at this time to determine whether or not students could complete a 30-item test with more appropriate items.

Identifying Unacceptable Items — The item response count obtained from the field testing of the first experimental editions of the frameworks has been analyzed for problems with individual test items. Consistent responses to the same distractor in an item were noted. The item was then analyzed to

check on the possibility of misleading content or two correct answers. Reading and mathematics content advisors from ESEA Title I field offices in the district and ESEA Title I coordinators of reading and mathematics programs in schools were called in to make judgments on those items. Items found to be unacceptable by these groups were revised to meet the necessary criteria. In some cases items were discovered that were responded to correctly, but did not actually measure performance on the objective with which they were identified. These were completely re-written with new content and formats.

Determining Reliability and Validity of Subdomain Scores —

Reliability: Coefficient alpha will be computed for each content subdomain within each testing framework. The procedure used for estimating the necessary components of variance (necessary statistics for estimating coefficient alpha) when matrix sampling is used is given by Shoemaker (1973).

Validity: The testing frameworks developed here are content valid in that the content of each framework is that agreed upon by all advisory panel members as representing what is or should be taught by teachers to Title I students. Each framework demonstrates additionally construct validity because the associated items are operational definitions of constructs defined by the content map.

Try Out and Use of Tests

Field Testing of the Frameworks — Both FAIR and FAIM have already been field tested in the spring and summer of 1976, and January of 1977. Item analysis by computer and critiques on item content and test format by

teachers and program advisors for the two subject areas have been obtained.

The second and third editions incorporated changes in test directions, illustrations, distractors and any portions of item content that may have been misleading or unrepresentative of the objectives.

The field testing of these second and third experimental editions is still a test of the test. Therefore, the results will not be used to assess student achievement. Instead, they will be used to detect further needs for revision.

The groups involved in these uses of the frameworks are as follows:

1) all ESEA Title I students in grades 7 through 12
2) 2,000 sixth grade ESEA Title I students
3) 2,400 control students from grades 6 (1,200), 8 (1.200) in schools not having ESEA Title I programs

Each time the frameworks are administered, one-third of each group listed above will be given FAIR, one-third FAIM, and one-third FAIES. Selection of subgroups of students to receive the three frameworks is made randomly.

Norming of the Frameworks — Over a three-year period, program assessment through the continued parallel use of the Comprehensive Tests of Basic Skills (CTBS) will be compared with results from FAIR and FAIM. Overall CTBS results will be compared to the number of items answered correctly on the two frameworks in an attempt to extra t normative data from FAIR, FAIM. There will be an attempt to establish local norms based on the use of FAIES.

Revision of Items for the Final Edition of the Framework — During the course of the first year using FAIR, FAIM, and FAIES, a special type of revision process is taking place. Selected students and teachers involved in the secondary compensatory education programs are being asked to participate in workshops for item revision. The object is to alter items where possible to be more relevant to the interests, maturity level, and cultural backgrounds of the students in the programs.

Suggestions are being asked for on changes in illustrations, contents of paragraphs for comprehension items, contents of graphs and word problems. It is hoped that such revisions will draw upon current interests of the students, and add relevance and humor to the items, allowing students to identify with the contents of the subtests.

Effects of Classroom Environment on Achievement — Test results obtained from the fall 1977 testing will be analyzed to determine the objectives on which students can and cannot perform.

Schools with poor results and schools with very good results (in both experimental and control groups) will be identified. Classrooms from these groups will then be randomly selected for a study of those attributes that comprise the instructional program in that setting. Such factors as management systems used, content covered and personnel used will be noted. Observations will be made of what students, teachers, and administrators do while in the classroom setting. These evaluations will be conducted over a three-year period to determine what progressive effects result from converting to domain-referenced testing and its use over that period of time.

Applicability of Domain-Referenced Test Results — After test results are released to the teachers of the examinees in the fall, each teacher will

receive a questionnaire asking them to rate and describe the value of domain-referenced test results to their program. Of interest to this project will be effects of the achievement information on such things as classroom practices, teaching methods, student attainment and general program organization.

The same questionnaire will be administered to the teachers after the spring testing results are released.

Also included will be questions about the value of test results obtained from domain-referenced tests versus those obtained from norm-referenced tests.

Training of Teachers for Use of the Frameworks — Two types of inservice training will be necessary to implement effective use of the frameworks. These will involve the following:

1) Instructing teachers on the administration of the framework
2) Explaining the meaning of the test results, how they differ from norm-referenced test results, and how the results may be applied to the program to make instructional decisions.

Parent Inservice — The parents of examinees will be offered information on the characteristics and intent of the domain-referenced tests. Differences from norm-referenced tests will be discussed and test results from use of the frameworks will be explained.

References to the Literature

Millman, Jason. "Passing Scores and Test Lengths for Domain-Referenced Measures"; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972). ERIC Report number; ED 065 555, U. D. Dept. of HEW, O. E.; Educational Resources Information Center, Washington, D. C. 20202.

Millman, Jason. "Criterion-Referenced Measurement" Evaluation in Education; W. James Popham, Editor. McCutchan Publishing Corporation; Berkeley, California 1974. p. 309

Roudbush, Glenn. "Normative Data from a CRT?" CRITERIA CTB Newsletter on Evaluation, no. 8. Published by CTB/McGraw-Hill; Monterey, California 1974.

Shoemaker, D. M. Principles and Procedures of Multiple Matrix Sampling. Cambridge, Mass.: Bollinger Publishing Company, 1973.

Shoemaker, David M. Toward a Framework for Achievement Testing, unpublished manuscript, second draft, 1974; through Southwest Regional Laboratory for Educational Research and Development.

Shoemaker, David M. "Toward a Framework for Achievement Testing," A Review of Educational Research, published by American Educational Research Association, 1975.